Lab #3
Power and sample size

In this lab, we are working with a data set from Alizadeh et al. at Stanford. In this study, the investigators were evaluating diffuse large B-cell lymphoma (DLBCL). Using expression profiling and hierarchical clustering (a topic that we will visit later in this class), they were able to identify 2 distinct forms of DLBCL that indicate different stages of B-cell differentiation. "One type expressed genes characteristic of germinal centre B cells ('germinal centre B-like DLBCL'); the second type expressed genes normally induced during in vitro activation of peripheral blood B cells ('activated B-like DLBCL')." They also found that the germinal centre B-like DLBCL patients had a better survival rate.

We will use this data set to evaluate the power and sample size in this experiment. We will also look for the necessary number of samples to appropriately power the study. First we will calculate the power and *n* required using a single gene calculation for illustration of the formula, then we will conduct a more multivariate summary that gives an idea of the power or *n* required for a specific percentage of genes/probes in the experiment. Remember that when we calculate these statistics for a microarray, we are dealing with more than a single variable, so general power formulas do not apply when attempting to summarize all genes/probes on an array.

The paper entitled "Distinct type of diffuse large B-cell lymphoma identified by gene expression profiling" can be found on the course website.

1.) Download the Eisen DLBCL data set and save as a text file (go to class web site or see syllabus for paper URL).

2.) Load into R, using read.table and arguments:

```
   header=T
   na.strings="NA"
   blank.lines.skip=F
   row.names=1
```

There are missing values in this data frame because we're working with cDNA data.

```
> dat <- read.table("/Users/stevendea/Desktop/JHU/Fall 2019/Gene Expression Data Analysis and Visualization/Labs/Lab 3/eisen.txt", header=T, na.strings="NA", blank.lines.skip=F, row.names=1)
```

3.) Get the class label file "eisenClasses.txt" from the class web site and read it into R. Use the header=T argument.

```
> ann <- read.table("/Users/stevendea/Desktop/JHU/Fall 2019/Gene Expression Data Analysis and Visualization/Labs/Lab 3/eisenClasses.txt", header=T)
```

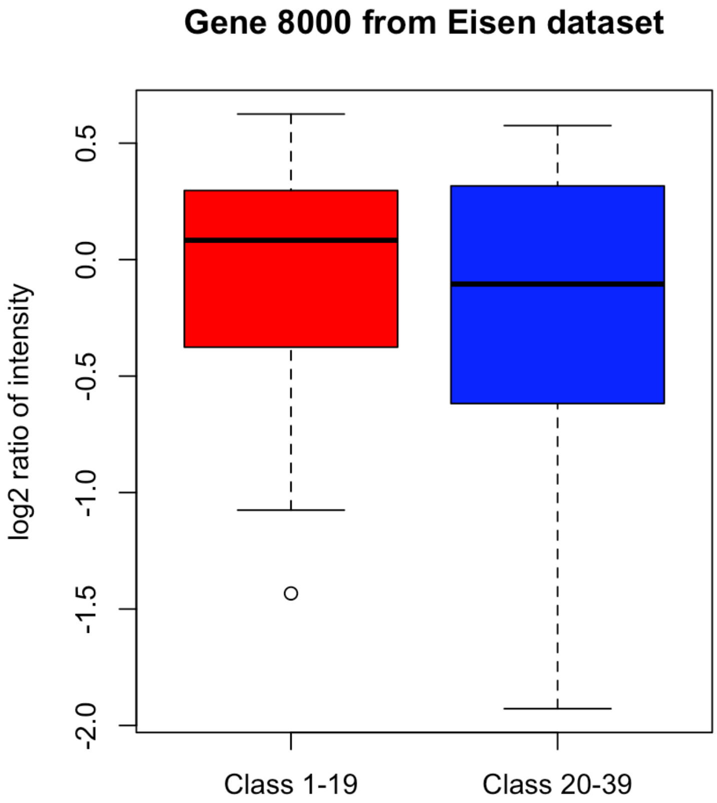
4.) Subset the data frame with the class labels and look at the positions so you know where one class ends and the other begins. Remember that 'subset' means to re-index (i.e. reorder) the column headers. If you look at the original column name order with dimnames(dat)[[2]] both before and after you reorder them, you will see what this has done.

```
> dat<-dat[,cl] #remove column that is not present in ann table
```

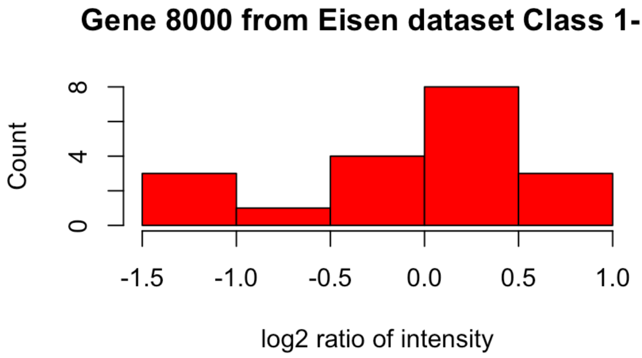
5.) Pick a gene, remove cells that have "NAs", and plot the values for both classes with a:

- boxplot (use the argument col=c(“red”,”blue”) to color separate boxes)
- histogram (this should have 2 separate histogram plots on 1 page; use the par(mfrow=c(2,1)) function prior to plotting the first).

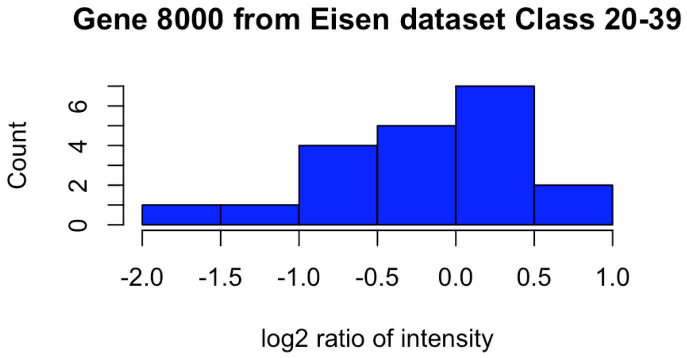
```
> x <- as.numeric(dat[8000,gc]) #search for gene 8000 in both classes
> y <- as.numeric(dat[8000,act])
> x <- x[!is.na(x)] #remove NA values
> y <- y[!is.na(y)]
> boxplot(x,y , col=c("red","blue"), names=c("Class 1-19", "Class 20-39"), main ="Gene 8000 from Eisen dataset", ylab="log2 ratio of intensity") #plot data found for gene 8000 in each class cohort
```



```
> par(mfrow=c(2,1)) #display 2 graphs per page
> hist(x, col="red", main ="Gene 8000 from Eisen dataset Class 1-19", xlab="log2
ratio of intensity", ylab="Count")
> hist(y, col="blue", main ="Gene 8000 from Eisen dataset Class 20-39",
xlab="log2 ratio of intensity", ylab="Count")
```



Gene 8000 from Eisen dataset Class 1-19



Gene 8000 from Eisen dataset Class 20-39

Color each class something different in the boxplot and histogram.

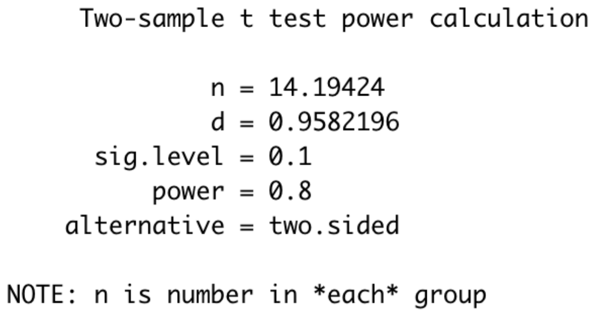
6.) Calculate the pooled variance as coded in the lecture notes, and calculate the minimum sample size necessary to detect a 1.5 fold difference (at 80% power and 99% confidence).

Pooled Variance:

```
> nx <- length(x)
> ny <- length(y)
> pool.var <-(((nx-1)*var(x)) + ((ny-1)*var(y)))/(nx+ny-2)
> pool.var
[1] 0.3726713
```

Min. sample size:

```
> dif.1.5fold <- log2(1.5)/sqrt(pool.var)
> pl.ss1.5 <- pwr.t.test(d=dif.1.5fold,sig.level=.1,power=0.8,type="two.sample")
```



```
     Two-sample t test power calculation 

              n = 14.19424
              d = 0.9582196
      sig.level = 0.1
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

7.) Now calculate the sample size required for the same gene selected in #5 using the empirically determined delta between the two groups, assuming 99% confidence and 80% power.

Since the gene I selected was 8000, same as coded in the lecture, I will select a different gene from #5 and calculate.

```
> x <- as.numeric(dat[4455,gc]) #search for gene 4455 in both classes
> y <- as.numeric(dat[4455,act])
> x <- x[!is.na(x)] #remove NA values
> y <- y[!is.na(y)]
> nx <- length(x)
> ny <- length(y)
> pool.var <-(((nx-1)*var(x)) + ((ny-1)*var(y)))/(nx+ny-2)
> pool.var
[1] 2.550484

> dif
[1] 0.1867293
```

Min. sample size:

```
> pl.ss <- pwr.t.test(d=dif.1.5fold,sig.level=.1,power=0.8,type="two.sample")
```

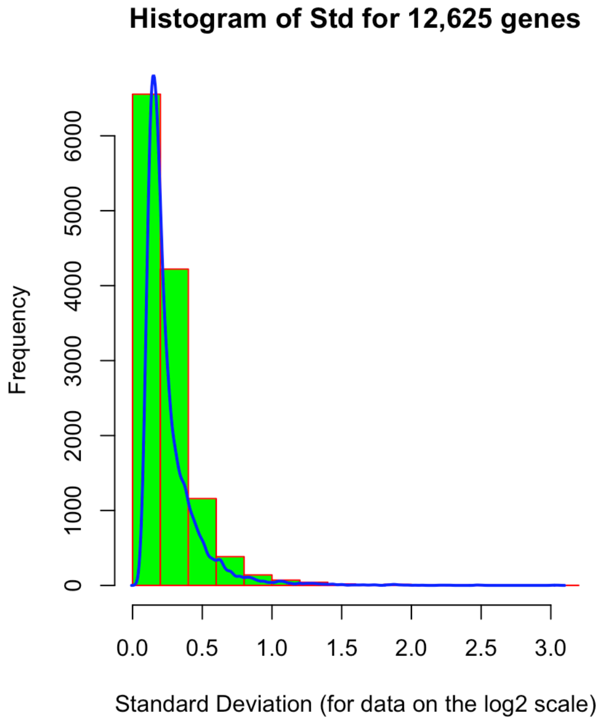
```
     Two-sample t test power calculation 

              n = 355.288
              d = 0.1867293
      sig.level = 0.1
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

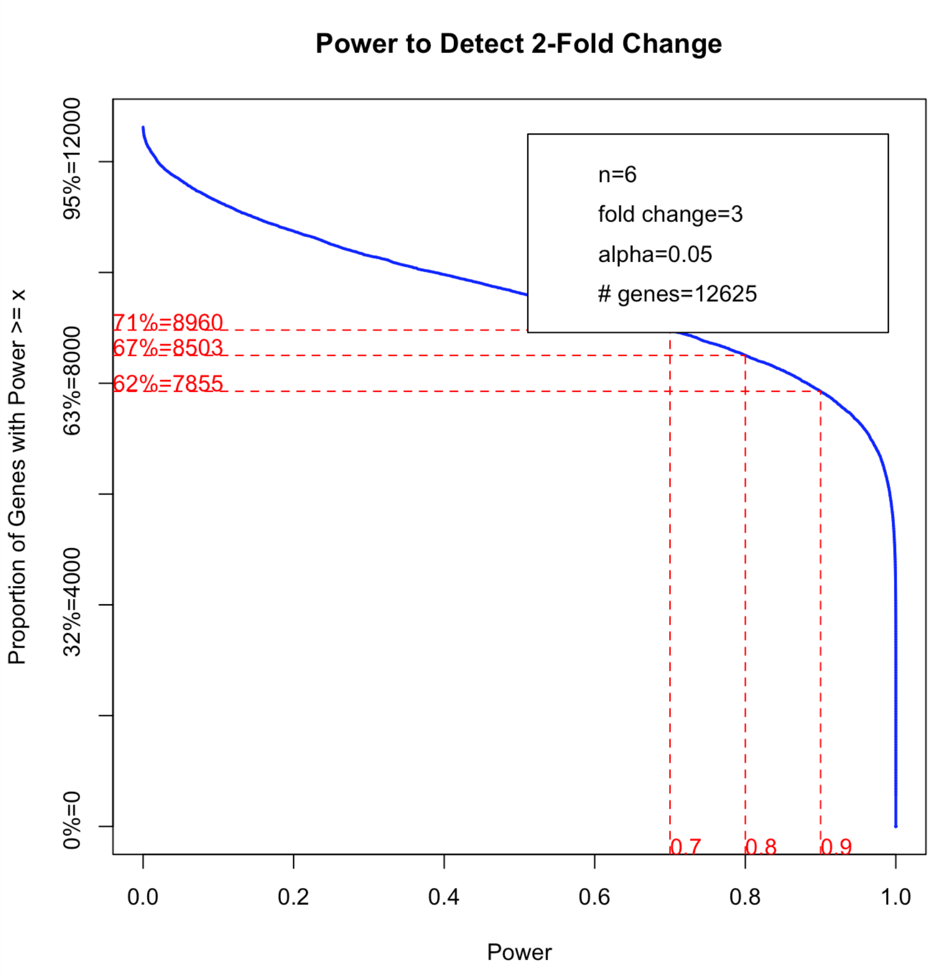
8.) Now load the ssize and gdata libraries, calculate the standard deviation for each gene in the matrix (Hint: use the na.rm=T argument), and plot a histogram of the standard deviations. Label the plot accordingly.

```
> library(ssize)
> library(gdata)
> data(exp.sd)
> hist(exp.sd, n=20, col="green", border ="red", main="", xlab="Standard Deviation (for data on the log2 scale)")
> lines(dens$x, dens$y*par("usr")[4]/max(dens$y),col="blue", lwd=2)
> title("Histogram of Std for 12,625 genes")
```



9.) Calculate and plot a proportion of genes vs. sample size graph to get an idea of the number of genes that have an adequate sample size for confidence=95%, effect size=3 (log2 transform for the function), and power=80%.

```
>n=6; fold.change=3.0; power=0.8; sig.level=0.05;
>all.power <- pow(sd=exp.sd, n=n, delta=log2(fold.change), sig.level=sig.level)
power.plot(all.power, lwd=2, col="blue")
>xmax <- par("usr")[2]-0.05
>ymax <- par("usr")[4]-0.05
>legend(x=xmax, y=ymax, legend= strsplit( paste("n=",n,",", "fold
change=",fold.change,",", "alpha=", sig.level, ",", "# genes=",length(exp.sd),
sep=''), "," )[[1]], xjust=1, yjust=1, cex=1.0)
title("Power to Detect 2-Fold Change")
```


Power to Detect 2-Fold Change
n=6
fold change=3
alpha=0.05
# genes=12625
71%=8960
67%=8503
62%=7855
Proportion of Genes with Power >= x
95%=12000
63%=8000
32%=4000
0%=0
Power